Optimization

Lectured by Dr Dante Kalise

Typed by Aris Zhu Yi Qing

April 27, 2022

Contents

Chapter 1

Mathematical Preliminaries

1.1 Topological Concepts

Definition 1. The **open ball** with center $c \in \mathbb{R}^n$ and radius r is

$$B(c,r) = \left\{\mathbf{x} : \|\mathbf{x} - c\| < r\right\}.$$

Similarly, the **closed ball** with center c and radius r is

$$B[c,r] = \left\{\mathbf{x} : \|\mathbf{x} - c\| \leq r\right\}.$$

Definition 2. Given a set $U \subseteq \mathbb{R}^n$, a point $\mathbf{c} \in U$ is called an **interior point** of U if $\exists r > 0$ for which $B(\mathbf{c}, r) \subseteq U$. The set of all interior points of a given set U is called the interior of the set and is denoted by

$$\text{int}(U) = \left\{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \text{ for some } r > 0\right\}.$$

Definition 3. Given a set $U \subseteq \mathbb{R}^n$, a **boundary point** of U is a vector $\mathbf{x} \in \mathbb{R}^n$ satisfying that any neighbourhood of $\mathbf{x}$ contains at least one point in U and at least one point in its completement U^c. We denote

$$\text{bd}(U) = \text{The set of all boundary points of a set } U.$$

Definition 4. The **closure** of a set $U \subseteq \mathbb{R}^n$ is the smallest closed set containing U, denoted by $\text{cl}(U)$ with

$$\text{cl}(U) = U \cup \text{bd}(U).$$

Definition 5. A set $U \subseteq \mathbb{R}^n$ is called **bounded** if $\exists M > 0$ for which $U \subseteq B(0, M)$.

Definition 6. A set $U \subseteq \mathbb{R}^n$ is called **compact** if it is closed and bounded.

1.2 Multi-variable Calculus

Definition 7. The **directional derivative** of a scalar function f w.r.t. $\mathbf{d}$ at a point $\mathbf{x}$ is denoted as

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d}$$

Theorem 8. Given the general quadratic functions of the form

$$f(\mathbf{w}) = \mathbf{w}^T A \mathbf{w} + \mathbf{b}^T \mathbf{w} + \gamma$$

we have

$$\nabla f(\mathbf{w}) = (A^T + A)\mathbf{w} + \mathbf{b}, \qquad \nabla^2 f(\mathbf{w}) = A + A^T.$$

If A is symmetric, then

$$\nabla f(\mathbf{w}) = 2A\mathbf{w} + \mathbf{b}, \qquad \nabla^2 f(\mathbf{w}) = 2A.$$

1.3 Positive Definiteness of Matrix

Proposition 9. Let A be a positive definite (semidefinite) matrix, then

- the diagonal elements of A are positive (nonnegative)
- $\mathrm{Tr}(A)$ and $\det(A)$ are positive (nonnegative)

(Test 1) Theorem 10. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then

- A is positive definite (semidefinite) iff all its eigenvalues are positive (nonnegative).
- A is indefinte iff it has at least one positive eigenvalue and at least one negative eigenvalue.

Definition 11. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then

- A is **diagonally dominant** if

$$|A_{ii}| \geq \sum_{j\neq i} |A_{ij}| \quad \forall i = 1, 2, \ldots, n$$

- A is **strictly diagonally dominant** if

$$|A_{ii}| > \sum_{j\neq i} |A_{ij}| \quad \forall i = 1, 2, \ldots, n$$

(Test 2) Theorem 12. If $A \in \mathbb{R}^{n \times n}$ is symmetric, diagonally dominant with positive (nonnegative) diagonal elements, then A is positive definite (semidefinite).

Chapter 2

Unconstrained Optimization

2.1 Optimums

Definition 13. Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$, then $\forall \mathbf{x} \in S$,

$$\mathbf{x}^* \in S \text{ is a } \textbf{global minimum} \text{ point of } f \text{ over } S \text{ if } f(\mathbf{x}) \geq f(\mathbf{x}^*),$$

$$\mathbf{x}^* \in S \text{ is a } \textbf{strict global minimum} \text{ point of } f \text{ over } S \text{ if } f(\mathbf{x}) > f(\mathbf{x}^*),$$

and similar definitions for maximum.

Definition 14. Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$, $\mathbf{x}^* \in S$ is a **local minimum** of f over S if $\exists r > 0$ s.t. $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for any $\mathbf{x} \in S \cap B(\mathbf{x}^*, r)$. Similar definitions for **strict local minimum** and maximum.

Definition 15. Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\mathbf{x}^* \in \text{int}(U)$ and that all the partial derivatives of f are defined at $\mathbf{x}^*$, then $\mathbf{x}^*$ is called a **stationary point** of f if $\nabla f(\mathbf{x}^*) = 0$.

2.2 Second-order Optimality Conditions

Theorem 16. Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point, then

- $\mathbf{x}^*$ is a local minimum point $\iff \nabla^2 f(\mathbf{x}^*) \succeq 0$.
- $\mathbf{x}^*$ is a strict local minimum point $\iff \nabla^2 f(\mathbf{x}^*) \succ 0$.
- similar necessary and sufficient conditions for (strict) local maximum point

Definition 17. Let $f : U \to \mathbb{R}$ be a continuously differentiable function defined on an open set $U \subseteq \mathbb{R}^n$. A stationary point $\mathbf{x}^* \in U$ is called a **saddle point** of f over U if it is neither a local minimum nor a local maximum point of f over U.

Theorem 18. Let $f : U \to \mathbb{R}$ be a continuously differentiable function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point. Then

$$\nabla^2 f(\mathbf{x}^*) \text{ is an indefinite matrix} \Longrightarrow \mathbf{x}^* \text{ is a saddle point of } f \text{ over } U.$$

2.3 Attainment of Minimal/Maximal Points

(Weierstrass') Theorem 19. Let f be a continuous function defined over a nonempty conpact set $C \subseteq \mathbb{R}^n$. Then $\exists$ a global minimum point of f over C and a global maximum point of f over C.

Definition 20. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function over $\mathbb{R}^n$. f is called **coercive** if

$$\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = \infty$$

Theorem 21. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous and coercive function and let $S \subseteq \mathbb{R}^n$ be a nonempty closed set. Then f attains a global minimum point on S.

2.4 Global Optimality Conditions

Theorem 22. Let f be a twice continuously differentiable function defined over $\mathbb{R}^n$. Let $\mathbf{x}^* \in \mathbb{R}^n$ be a stationary point of f. Then

$$\nabla^2 f(\mathbf{x}) \succeq 0 \; \forall \mathbf{x} \in \mathbb{R}^n \Longrightarrow \mathbf{x}^* \text{ is a global minimum point of } f.$$

Proposition 23. Let $f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} + 2\mathbf{b}^T \mathbf{x} + c$, with $A \in \mathbb{R}^{n \times n}$ symmetric, then

1. $\mathbf{x}$ is a stationary point of f iff $A\mathbf{x} = -\mathbf{b}$.
2. if $A \succeq 0$, then $\mathbf{x}$ is a global minimum point of f iff $A\mathbf{x} = -\mathbf{b}$.
3. if $A \succ 0$, then $\mathbf{x} = -A^{-1}\mathbf{b}$ is a strict global minimum point of f.

Chapter 3

Linear Least Squares

3.1 Problem Formulation

Consider the linear system

$$S\mathbf{x} \approx \mathbf{b}, \quad (S \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, m > n)$$

To solve the above system, the usual approach is to transform it to become

$$\min_{\mathbf{x}} \|S\mathbf{x} - \mathbf{b}\|^2 \iff \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \mathbf{x}^T S^T S\mathbf{x} - 2\mathbf{b}^T S\mathbf{x} + \|\mathbf{b}\|^2 \right\}.$$

Note that $\nabla^2 f(\mathbf{x}) = 2S^T S \succeq 0$ since $\mathbf{x}^T S^T S\mathbf{x} = (S\mathbf{x})^T (S\mathbf{x}) = \|S\mathbf{x}\|^2 \geq 0$. Therefore, the unique optimal solution $\mathbf{x}_{\text{LS}}$ is the solution $\nabla f(\mathbf{x}) = 0$, namely

$$(S^T S)\mathbf{x}_{\text{LS}} = S^T\mathbf{b} \Longrightarrow \mathbf{x}_{\text{LS}} = \left(S^T S\right)^{-1} S^T\mathbf{b}.$$

3.2 Data Fitting

1. For dataset $(\mathbf{s}_i, b_i)$ where $\mathbf{s}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, we could transform to problem

$$\min_{\mathbf{x}} \sum_{i=1}^{m} \left(\mathbf{s}_i^T\mathbf{x} - b_i\right)^2 \Longrightarrow \min_{\mathbf{x}} \|S\mathbf{x} - \mathbf{b}\|^2$$

2. For polynomial fitting, given a set of points $\mathbb{R}^2 : (u_i, y_i)$, the associated linear system is

$$\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}$$

3.3 Regularized Least Squares

A Regularized Least Square problem is formulated as

$$\min_{\mathbf{x}} \|S\mathbf{x} - \mathbf{b}\|^2 + \lambda R(\mathbf{x}),$$

where λ is the regularization parameter and $R(\cdot)$ is the regularization function (also called a *penalty* function). A common choice is a quadratic regularization function:

$$\min_{\mathbf{x}} \|S\mathbf{x} - \mathbf{b}\|^2 + \lambda \|D\mathbf{x}\|^2$$

with its optimal solution being

$$\mathbf{x}_{\text{RLS}} = \left(S^T S + \lambda D^T D\right)^{-1} S^T \mathbf{b}$$

since $\nabla f = 2S^T S\mathbf{x} - 2S^T\mathbf{b} + 2\lambda D^T D\mathbf{x} = 0$.

3.4 Denoising

Suppose a noisy measurement of a signal $\mathbf{x} \in \mathbb{R}^n$ is given

$$\mathbf{b} = \mathbf{x} + \mathbf{w}$$

where $\mathbf{x}$ is the "true" unknown signal, $\mathbf{w}$ is the unknown noise and $\mathbf{b}$ is the (known) measures vector. We could define

$$R(\mathbf{x}) = \|L\mathbf{x}\|^2, \text{ where } L = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

as the regularization function to penalize any sudden variations in signal. The RLS is thus

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|^2 + \lambda \|L\mathbf{x}\|^2$$

with its direct solution being

$$\mathbf{x}_{\text{RLS}}(\lambda) = \left(I + \lambda L^T L\right)^{-1} \mathbf{b}.$$

Chapter 4

The Gradient Method

4.1 Descent Direction

Definition 24. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. A vector $\mathbf{0} \neq \mathbf{d} \in \mathbb{R}^n$ is called a **descent direction** of f at $\mathbf{x}$ if

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d} < 0.$$

Example 25. The descent direction can be $\mathbf{d} = -\nabla f(\mathbf{x})$, since as long as $\nabla f(\mathbf{x}) \neq 0$ ($\mathbf{x}$ is a non-stationary point), we have

$$f'(\mathbf{x}; -\nabla f(\mathbf{x})) = -\nabla f(\mathbf{x})^T f(\mathbf{x}) = -\left\|\nabla f(\mathbf{x})\right\|^2 < 0.$$

Lemma 26. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Let $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$, then

$$\exists \epsilon > 0 \text{ s.t. } \forall t \in (0, \epsilon], f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x}).$$

Lemma 27. Let f be a continuously differentiable function and $\mathbf{x} \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(\mathbf{x}) \neq 0$), then the optimal solution of

$$\min_{\mathbf{d}} \left\{ f'(\mathbf{x}; \mathbf{d} : \|\mathbf{d}\| = 1 \right\}$$

is $\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{\left\|\nabla f(\mathbf{x})\right\|}$.

Lemma 28. Let $\left\{\mathbf{x}^k\right\}_{k>0}$ be the sequence generated by the gradient descent method with *exact* line search for solving a problem of minimizing a continuously differentiable function f. Then $\forall k = 0, 1, 2, \ldots$,

$$(\mathbf{x}^{k+2} - \mathbf{x}^{k+1})^T (\mathbf{x}^{k+1} - \mathbf{x}^k) = 0.$$

4.2 Stepsize Selection Rules

Finding the right $t^k \in \mathbb{R}^n$, called the **stepsize**, is referred in the literature as **line search**.

1. Constant stepsize: $t^k = \bar{t}\ \forall k$.
2. Exact stepsize: t^k is a minimizer of f along the ray $\mathbf{x}_t^k \mathbf{d}^k$:
$$t^k \in \underset{t \geq 0}{\operatorname{argmin}} f(\mathbf{x}^k + t\mathbf{d}^k)$$
3. Backtracking (Armijo rule): let $s > 0, \alpha \in (0,1), \beta \in (0,1)$, and initial stepsize $t^k = s$, while
$$f(\mathbf{x}^k) - f(\mathbf{x}^k + t^k\mathbf{d}^k) < -\alpha t^k \nabla f(\mathbf{x}^k)^T \mathbf{d}^k$$
set $t^k := \beta t^k$, iterating until achieving the **sufficient decrease property**
$$f(\mathbf{x}^k) - f(\mathbf{x}^k + t^k\mathbf{d}^k) \geq -\alpha t^k \nabla f(\mathbf{x}^k)^T \mathbf{d}^k.$$

4.3 Convergence

Definition 29. Let f be a continuously differentiable function over $\mathbb{R}^n$. We say that f has a **Lipschitz gradient** if

$$\exists L \geq 0 \text{ s.t. } \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

L is called the **Lipschitz constant**.

Comments:

- The class of functions with Lipschitz gradient with constant L is denoted as $C_L^{1,1}(\mathbb{R}^n)$ or just $C_L^{1,1}$. When L is irrelevant, we simply denote the class by $C^{1,1}$.
- If ∇f is Lipschitz with constant L, then it is also Lipschitz with constant $L'\ \forall L' \geq L$.
- Linear functions: Given $a \in \mathbb{R}^n$, the function $f(\mathbf{x}) = a^T\mathbf{x}$ is in $C_0^{1,1}$.
- Quadratic functions: Let $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$, then the function $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ is $C_{2\|A\|_2}^{1,1}$.

Theorem 30. Let f be a continuously differentiable function over $\mathbb{R}^n$. Then

$$f \in C_L^{1,1}(\mathbb{R}^n) \iff \left\|\nabla^2 f(\mathbf{x})\right\| \leq L\ \forall \mathbf{x} \in \mathbb{R}^n.$$

(Sufficient decrease of the gradient method) Lemma 31. Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$,
- exact line search
- backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, and $\beta \in (0,1)$,

then

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M \left\| \nabla f(\mathbf{x}^k) \right\|^2$$

where

$$M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right) & \text{constant stepsize} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} & \text{backtracking} \end{cases}$$

(Convergence of the gradient method) Theorem 32. Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and is bounded below over $\mathbb{R}^n$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$,
- exact line search
- backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, and $\beta \in (0,1)$,

then

1. $\forall k, f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ unless $\nabla f(\mathbf{x}^k) = 0$.
2. $\nabla f(\mathbf{x}^k) \to 0$ as $k \to \infty$.

4.4 Condition Number and Convergence for Quadratic Function

Definition 33. Let $A \in \mathbb{R}^{n\times n}$ be positive definite, Then the **condition number** of A is

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the largest and smallest eigenvalues respectively.

(Kantorovich inequality)Lemma 34. Let $A \in \mathbb{R}^{n\times n}$ be positive definite. Then

$$\forall \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n,\ \frac{\left(\mathbf{x}^T\mathbf{x}\right)^2}{(\mathbf{x}^T A\mathbf{x})(\mathbf{x}^T A^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{\left(\lambda_{\max}(A)+\lambda_{\min}(A)\right)^2}.$$

(Convergence for quadratic function) Theorem 35. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}^T A\mathbf{x}) \quad (A \succ 0),$$

then $\forall k = 0, 1, \dots,$

$$f(\mathbf{x}^{k+1}) \leq \left(\frac{\lambda_{\max}(A)-\lambda_{\min}(A)}{\lambda_{\max}(A)+\lambda_{\min}(A)}\right)^2 f(\mathbf{x}^k) = \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^2 f(\mathbf{x}^k).$$

4.5 Scaled Gradient Method

A way to mitigate the slow convergence due to poor conditioning of the Hessian is to formulate a rescaled version of the problem. From the minimization problem

$$\min\left\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\right\}$$

we introduce a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ to make the linear change of variables $\mathbf{x} = S\mathbf{y}$ and obtain the equivalent problem

$$\min\left\{g(\mathbf{y}) \equiv f(S\mathbf{y}) : \mathbf{y} \in \mathbb{R}^n\right\}$$

Since $\nabla g(\mathbf{y}) = S^T\nabla f(S\mathbf{y}) = S^T\nabla f(\mathbf{x})$, the gradient method for the rescaled problem reads

$$\mathbf{y}^{k+1} = \mathbf{y}^k - t^k S^T \nabla f(S\mathbf{y}^k).$$

Multiplying both sides by S, with $\mathbf{x}^k = S\mathbf{y}^k$, and define $D = SS^T$, we have

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k D\nabla f(\mathbf{x}^k).$$

Since $D \succ 0$, so

$$f'(\mathbf{x}^k; -D\nabla f(\mathbf{x}^k)) = -\nabla f(\mathbf{x}^k)^T D\nabla f(\mathbf{x}^k) < 0.$$

A well-known choice for D^k is to pick $D^k = \left(\nabla^2 f(\mathbf{x}^k)\right)^{-1}$ (Newton's method). Another alternative is to use a diagonal scaling, e.g.

$$\left(D^k\right)_{ii} = \left(\frac{\partial^2 f(\mathbf{x}^k)}{\partial x_i^2}\right)^{-1}$$

4.6 The Kaczmarz Algorithm

The *Kaczmarz Algorithm* solves the linear system

$$A\mathbf{x} = \mathbf{b}$$

by iterating projections along the i-th row of the matrix A, denoted by $\mathbf{a}_i^T$:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{b_i - \mathbf{a}_i^T \mathbf{x}^k}{\|\mathbf{a}_i\|^2} \mathbf{a}_i$$

In the original Kaczmarz algorithm, the i-th row is chosen periodically by cycling through all rows. If chooses i-th row randomly, we can show that the algorithm converges exponentially, and this is known as *randomized Kaczmarz Algorithm.*

The algorithm works because the problem of solving the linear system $A\mathbf{x} = \mathbf{b}$ could be formulated as an optimization problem

$$\min_{\mathbf{x}} \frac{1}{2m} \|A\mathbf{x} - \mathbf{b}\|^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(\mathbf{a}_i^T \mathbf{x} - \mathbf{b}_i\right)^2$$

for which the gradient descent method could be constructed as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{t}{m} A^T (A\mathbf{x} - \mathbf{b})$$

but the problem could also be formulated as

$$\min_{\mathbf{x}} \frac{1}{2m} \|A\mathbf{x} - \mathbf{b}\|^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(\mathbf{a}_i^T \mathbf{x} - b_i\right)^2 = \frac{1}{2} \mathbb{E}_i [\mathbf{a}_i^T \mathbf{x} - b_i],$$

which can then be translated to the action of randomly picking a row of A, becoming

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{t}{m} (\mathbf{a}_i^T \mathbf{x} - b_i) \mathbf{a}_i$$

4.7 Stochastic Gradient Descent

Theorem 36. Assuming that

- The cost $g(\mathbf{x})$ is such that
$$\left\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\right\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \text{and} \quad \nabla^2 g(\mathbf{x}) \succeq \mu I.$$
- The sample gradient $\nabla Q_i(\mathbf{x}^k)$ is an unbiased estimate of $\nabla g(\mathbf{x}^k)$.
- $$\forall \mathbf{x}, \mathbb{E}_i \left[\left\|Q_i(\mathbf{x})\right\|^2\right] \leq \sigma^2 + c\left\|\nabla g(\mathbf{x})\right\|^2.$$

Then if $t^k \equiv t \leq \frac{1}{Lc}$, then SGD achieves
$$\mathbb{E}\left[g(\mathbf{x}^k) - g(\mathbf{x}^*)\right] \leq \frac{tL\sigma^2}{2\mu} + (1 - t\mu)^k (g(\mathbf{x}^0) - g(\mathbf{x}^*)).$$

Comments

1. Fast (linear) convergence during the first iterations.
2. Convergence to a neighbourhood of $\mathbf{x}^*$, without further progress.
3. If gradient computation is noiseless ($\sigma = 0$), then linear convergence to optimal point.
4. A smaller stepsize t yield better converging points.

Definition 37. The **batch gradient descent** algorithm is defined as
$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla g(\mathbf{x}^k) = \mathbf{x}^k - \frac{t^k}{|K|} \sum_{i \in K} \nabla Q_i(\mathbf{x}^k),$$
where K denotes a set of p randomly selected datapoints.

Chapter 5

Convexity

5.1 Convex Sets

Definition 38. A set $C \subseteq \mathbb{R}^n$ is called **convex** if

$$\forall \mathbf{x}, \mathbf{y} \in C \text{ and } \lambda \in [0,1], \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in C.$$

Equivalently, for any $\mathbf{x}, \mathbf{y} \in C$, the line segment $[\mathbf{x}, \mathbf{y}]$ is also in C.

Example 39. Very important convex sets

- A line in $\mathbb{R}^n$ is a set of the form

  $$L = \{\mathbf{z} + t\mathbf{d} : t \in \mathbb{R}\},$$

  where $\mathbf{z}, \mathbf{d} \in \mathbb{R}^n$ and $\mathbf{d} \neq \mathbf{0}$.

- $[\mathbf{x}, \mathbf{y}], (\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n (\mathbf{x} \neq \mathbf{y})$, $\emptyset$, and $\mathbb{R}^n$.

- A **hyperplane** is a set of the form

  $$H = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x} = b\right\} \quad (\mathbf{a} \in \mathbb{R} \backslash \{\mathbf{0}\}, b \in \mathbb{R})$$

- The associated **half space** is the set

  $$H^- = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x} \leq b\right\}.$$

- The open ball $B(\mathbf{c}, r)$ and the closed ball $B[\mathbf{c}, r]$.

- The **ellipsoid** is a set of the form

  $$E = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T Q \mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c \leq 0\right\}$$

  where $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Lemma 40. Let $C_i \subseteq \mathbb{R}^n$ be a convex set for any $i \in I$, where I is an index set (possibly infinite), then $\bigcap_{i\in I} C_i$ is convex.

<u>**Comments**</u>: A direct consequence of the above is that convex polytopes of the form

$$P = (\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \leq \mathbf{b}),$$

are convex since they are generated as the intersection of m half-spaces $\mathbf{a}_i^T\mathbf{x} \leq b_i$.

Theorem 41. Several important algebraic properties of convex sets:

1. Let $C_1, C_2, \ldots, C_k \subseteq \mathbb{R}^n$ be convex sets and let $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$, then the set $\mu_1 C_1 + \mu_2 C_2 + \cdots + \mu_k C_k$ is convex.
2. Let $C_i \subseteq \mathbb{R}^{k_i}$, $i = 1, \ldots, m$ be convex sets, then the cartesian product
$$C_1 \times C_2 \times \cdots \times C_m = \left\{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) : \mathbf{x}_i \in C_i, i = 1, 2, \ldots, m\right\}$$
is convex.
3. Let $M \subseteq \mathbb{R}^n$ be a convex set and let $A \in \mathbb{R}^{m\times n}$, then the set
$$A(M) = \{A\mathbf{x} : \mathbf{x} \in M\}$$
is convex.
4. Let $D \subseteq \mathbb{R}^m$ be convex and let $A \in \mathbb{R}^{m\times n}$, then the set
$$A^{-1}(D) = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \in D\}$$
is convex.

5.2 Convex Hull

Definition 42. Given m points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^n$, a **convex combination** of these m points is a vector of the form

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_m\mathbf{x}_m$$

where $\lambda_i \in \mathbb{R}_+$ for $i = 1, 2, \ldots, m$ and satisfy $\sum_{i=1}^m \lambda_i = 1$ ($\boldsymbol{\lambda} \in \Delta_m$).

Theorem 43. Let $C \subseteq \mathbb{R}^n$ be a convex set and let $\mathbf{x}_i \in C$ for $i = 1, 2, \ldots, m$. Then for any $\boldsymbol{\lambda} \in \Delta_m$, the relation

$$\sum_{i=1}^m \lambda_i\mathbf{x}_i \in C$$

holds.

Definition 44. Let $S \subseteq \mathbb{R}^n$. The **convex hull** of S, denoted by $\operatorname{conv}(S)$, is the set comprising all the convex combinations of vectors from S:

$$\operatorname{conv}(S) = \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{x}_i : \mathbf{x}_i, \mathbf{x}_2, \ldots, \mathbf{x}_k \in S, \boldsymbol{\lambda} \in \Delta_k \right\}$$

<u>**Comment**</u>: $\operatorname{conv}(S)$ is the "smallest" convex set containing S.

Theorem 45. Let $S \subseteq \mathbb{R}^n$ and let $\mathbf{x} \in \operatorname{conv}(S)$. Then

$$\exists \boldsymbol{\lambda} \in \Delta_{n+1}, \exists \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1} \in S \text{ s.t. } \mathbf{x} = \sum_{i=1}^{n+1} \lambda_i \mathbf{x}_i.$$

Example 46. For $n = 2$, consider the four vectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix},$$

and let $\mathbf{x} \in \operatorname{conv}(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\})$ be given by

$$\mathbf{x} = \frac{1}{8}\mathbf{x}_1 + \frac{1}{4}\mathbf{x}_2 + \frac{1}{2}\mathbf{x}_3 + \frac{1}{8}\mathbf{x}_4 = \begin{pmatrix} \frac{13}{8} \\ \frac{11}{8} \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{\lambda} = \begin{pmatrix} 1/8 \\ 1/4 \\ 1/2 \\ 1/8 \end{pmatrix},$$

We can find out that

$$(\mathbf{x}_2 - \mathbf{x}_1) + (\mathbf{x}_3 - \mathbf{x}_1) - (\mathbf{x}_4 - \mathbf{x}_1) = 0 \quad \Longrightarrow \quad \boldsymbol{\mu} = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}.$$

Since we need to satisfy that $\forall i \in \{1, 2, 3, 4\}$, $\lambda_i + \alpha\mu_i \geq 0$, we need to compute

$$\epsilon = \min_{i:\mu_i<0} \left\{ -\frac{\lambda_i}{\mu_i} \right\}$$

so that $\lambda_j + \epsilon\mu_j = 0$ for $j \in \operatorname{argmin}_{i:\mu_i<0} \left\{ -\frac{\lambda_i}{\mu_i} \right\}$, thereby reducing the number of $\mathbf{x}_i$'s required for expressing $\mathbf{x}$. From the four inequalities, we can obtain that

$$\begin{cases} \alpha \leq 1/8 \\ \alpha \geq -1/4 \\ \alpha \geq -1/2 \\ \alpha \leq 1/8 \end{cases}$$

and $\epsilon = \frac{1}{8}$. Substituting $\alpha = \epsilon$, we can obtain that

$$\mathbf{x} = \frac{3}{8}\mathbf{x}_2 + \frac{5}{8}\mathbf{x}_3.$$

Definition 47. Let $S \subseteq \mathbb{R}^n$ be a convex set. A point $\mathbf{x} \in S$ is called an **extreme point** of S if $\nexists \mathbf{x}_1, \mathbf{x}_2 \in S(\mathbf{x}_1 \neq \mathbf{x}_2$ and $\lambda \in (0,1)$, s.t. $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. The set of extreme point is denoted by $\text{ext}(S)$.

Theorem 48. Let $S \subseteq \mathbb{R}^n$ be a compact convex set. Then

$$S = \text{conv}(\text{ext}(S)).$$

5.3 Convex Functions

Definition 49. A function $f : C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is called **convex** (or convex over C) if

$$\forall \mathbf{x}, \mathbf{y} \in C, \lambda \in [0,1], f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$$

Definition 50. A function $f : C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is called **strict convex** if

$$\forall \mathbf{x} \neq \mathbf{y} \in C, \lambda \in (0,1), f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$$

Definition 51. A function is called **concave** if $-f$ is convex. Similarly, f is called **strictly concave** if $-f$ is strictly convex.

Example 52. Several examples of convex functions:

- Affine functions: $f(\mathbf{x}) = a^T\mathbf{x} + b$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$, then
$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}).$$
- Norms: $g(\mathbf{x}) = \|\mathbf{x}\|$. Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$, then
$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \|\lambda\mathbf{x}\| + \big\|(1-\lambda)\mathbf{y}\big\| = \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})$$

(Jensen's Inequality) Theorem 53. Let $f : C \to \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^n$ is a convex set. Then $\forall \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in C$ and $\boldsymbol{\lambda} \in \Delta_k$,

$$f\left(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i).$$

5.4 First-order Characterization of Convex Functions

Theorem 54. Let $f : C \to \mathbb{R}$ be a continuously differentiable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then

$$f \text{ is convex over } C \quad \Longleftrightarrow \quad \forall \mathbf{x}, \mathbf{y} \in C, f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y})$$

An analogous result holds for strictly convex functions with a strict inequality.

Comment: For a convex function f defined on $\mathbb{R}^2$, the tangent plane at every point is always below f.

(Global optimality test for convex(concave) function) Theorem 55. Let f be a continuously differentiable function which is convex over a convex set $C \subseteq \mathbb{R}^n$. Then

$$\nabla f(\mathbf{x}^*) = 0 \text{ for some } \mathbf{x}^* \in C \quad \Longrightarrow \quad \mathbf{x}^* \text{ is the global minimizer of } f \text{ over } C.$$

This is the same for concave function being related to global maximizer.

(Convexity of quadratic function) Theorem 56. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the quadratic function given by $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ where $A \in \mathbb{R}^{n\times n}$ is symmetric, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then

$$f \text{ is (strictly) convex} \quad \Longleftrightarrow \quad A \succeq 0 (A \succ 0).$$

(Monotonicity of the gradient) Theorem 57. Suppose that f is a continuously differentiable function over a convex set $C \subseteq \mathbb{R}^n$, then

$$f \text{ is convex over } C \quad \Longleftrightarrow \quad \forall \mathbf{x}, \mathbf{y} \in C, (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq 0.$$

An analogous result holds for strictly convex functions with a strict inequality.

Proof. If f is convex, then

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$$

and

$$f(x) \geq f(y) + \nabla f(y) \cdot (x - y)$$

so that by adding the above inequalities, we obtain the result. □

5.5 Second-order Characterization of Convex Functions

Theorem 58. Let f be a twice continuously differentiable function over an open convex set $C \subseteq \mathbb{R}^n$. Then

$$f \text{ is convex over } C \quad \Longleftrightarrow \quad \forall \mathbf{x} \in C, \nabla^2 f(\mathbf{x}) \succeq 0$$

Example 59. Convexity of the log-sum-exp function

$$f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \cdots + e^{x_n}\right),\ \mathbf{x} \in \mathbb{R}^n.$$

The gradient is given by

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, \quad i = 1, 2, \ldots, n.$$

Therefore, the Hessian is computed as

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} -\frac{e^{x_i} e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2} & i \neq j \\ -\frac{e^{x_i} e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2} + \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} & i = j \end{cases}$$

We can thus write the Hessian matrix as

$$\nabla^2 f(\mathbf{x}) = \operatorname{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T, \quad \text{with} \quad \mathbf{w} = \left(\frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}\right)_{i=1}^n \in \Delta_n.$$

For any $\mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{i=1}^n w_i v_i^2 - \left(\mathbf{v}^T \mathbf{w}\right)^2 \geq 0,$$

since defining $s_i = \sqrt{w_i} v_i, t_i = \sqrt{w_i}$, we have

$$(\mathbf{v}^T \mathbf{w})^2 = \left(\mathbf{s}^T \mathbf{t}\right)^2 \leq \|\mathbf{s}\|^2 \|\mathbf{t}\|^2 = \left(\sum_{i=1}^n w_i v_i^2\right) \left(\sum_{i=1}^n w_i\right) = \sum_{i=1}^n w_i v_i^2.$$

Thus $\nabla^2 f(\mathbf{x}) \succeq 0$ and hence f is convex over $\mathbb{R}^n$.

5.6 More Results of Convex Function

Theorem 60. Let $f, f_1, f_2, \ldots, f_p$ be convex functions over a convex set $C \subseteq \mathbb{R}^n$.

- Let $\alpha \geq 0$, then αf is a convex function over C.
- The sum function $\sum_{i=1}^p f_i$ is convex over C.
- Let $A \in \mathbb{R}^{n \times m}$ and $\mathbf{b} \in \mathbb{R}^n$. Then the function $g(\mathbf{y}) = f(A\mathbf{y} + \mathbf{b})$ is convex over the convex set $D = \{\mathbf{y} \in \mathbb{R}^m : A\mathbf{y} + \mathbf{b} \in C\}$.
- Let $g : I \to \mathbb{R}$ be a nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume that the image of C under f is contained in I: $f(C) \subseteq I$, then the composition of g and f defined by $h(\mathbf{x}) \equiv g(f(\mathbf{x}))$ is convex over C.

(Point-wise maximum of convex functions) Theorem 61. Let $f_1, f_2, \ldots, f_p : C \to \mathbb{R}$ be p convex functions over the convex set $C \subseteq \mathbb{R}^n$, then the maximum function

$$f(\mathbf{x}) \equiv \max_{i=1,2,\ldots,p} \left\{ f_i(\mathbf{x}) \right\}$$

is convex over C.

Theorem 62. Let $f : C \times D \to \mathbb{R}$ be a convex function defined over the set $C \times D$ where $C \subseteq \mathbb{R}^m$ and $D \subseteq \mathbb{R}^n$ are convex sets. Let

$$g(\mathbf{x}) = \min_{\mathbf{y} \in D} f(\mathbf{x}, \mathbf{y}), \quad \mathbf{x} \in C$$

where we assume that the minimum is finite. Then g is convex over C.

Example 63. The distance function from a convex set $d_C(\mathbf{x}) \equiv \inf_{y \in C} \|\mathbf{x} - \mathbf{y}\|$.

Theorem 64. Let $f : C \to \mathbb{R}$ be a convex function defiend over a convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x}_0 \in \text{int}(C)$. Then $\exists \epsilon > 0, L > 0$ s.t. $B[\mathbf{x}_0, \epsilon] \subseteq C$ and

$$\forall \mathbf{x} \in B[\mathbf{x}_0, \epsilon], |f(\mathbf{x}) - f(\mathbf{x}_0)| \leq L \|\mathbf{x} - \mathbf{x}_0\| .$$

Theorem 65. Let $f : C \to \mathbb{R}$ be a convex function over the convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x} \in \text{int}(C)$. Then

$$\forall \mathbf{d} \neq \mathbf{0}, \exists f'(\mathbf{x}; \mathbf{d}).$$

Theorem 66. Let $f : C \to \mathbb{R}$ be convex and non-constant over the nonempty convex set $C \subseteq \mathbb{R}^n$. Then f does not attain a maximum at a point in $\text{int}(C)$.

Theorem 67. Let $f : C \to \mathbb{R}$ be convex over the nonempty convex and compact set $C \subseteq \mathbb{R}^n$. Then there exists at least one maximizer of f over C that is an extreme point of C.